The Erdős Institute Data Science Bootcamp Summer 2024
Executive Summary

Team QED:
Adnan Cihan Cakar, Matthew Gelvin, Cisil Karaguzel, Hatice Mutlu, Ming Zhang

Github: link

Overview:

The state-of-the-art language models have achieved human-level performance on many tasks but still face significant challenges in multi-step mathematical reasoning. Recent advancements in large language models (LLMs) have demonstrated exceptional capabilities across diverse tasks, including common-sense reasoning, question answering, and summarization. However, they struggle with tasks requiring quantitative reasoning, such as solving complex mathematical problems. Mathematics serves as a valuable testbed in machine learning for problem-solving abilities, highlighting the need for more robust models capable of multi-step reasoning.

Objective:

The primary goal of this project is to develop a customized LLM that can provide step-by-step solutions to math problems by fine-tuning a base LLM using a large mathematical dataset. The major priorities of the project are:

- Fine-tuning the model to handle mathematical notation and reasoning.
- Improving accuracy from the base model.

Training and Evaluation:

Training Dataset:

- The MetaMathQA dataset is used for training. This dataset is available on Hugging Face and contains 395k math questions augmented from the **training portions** of the GSM8K and MATH datasets:
    - The GSM8k dataset (training part only) contains grade school level math problems requiring 2 to 8 steps to solve.
    - The MATH dataset (training part only) includes more challenging problems from prestigious math competitions such as AMC 10, AMC 12, and AIME.

Model and Training Procedure:

- Starting with Unsloth's **llama-3-8b-bnb-4bit** model, a quantized 4-bit model using bitsandbytes, which allows for faster fine-tuning and reduced memory usage. It is a small model with 4.65B parameters.
- Adding LoRA adapters to update only 1-10% of parameters, allowing faster fine-tuning.
- Using the Alpaca prompt to format our dataset.
- Fine-tuning the base model with supervised instruction-response pairs using Hugging Face's SFFT trainer for one epoch.
- The model was trained with A100 GPU for 16 hours.

Model Testing:

- We performed a hyperparameter search on the temperature of the model, which controls the variance of the output.
- Performance was judged by autoregressively sampling a single low-temperature solution, extracting and checking if the final answer is correct.
- The evaluation datasets include **the test portions** of GSM8k and MATH, which are independent of the training dataset and ideal for assessing our model's generalizability.
- When computing the accuracy of our model, the evaluation functions (extracting, parsing, and comparing math quantities) developed by the Xwin-Math team were used.
- Solution generation with and without majority voting were both tested.

Results:

- *Zero-shot Accuracy:* We randomly selected 500 questions from each evaluation dataset and computed accuracies based on one generated solution for each question. Our model has relative improvements of 20% and 11% over the base model on the two datasets.

Model	GSM8k	Math
Base model (llama-3-8b-bnb-4bit)	29.4%	7.00%
Our model	35.2% +20%	7.80% +11%

- Majority vote: We evaluated our model's performance using majority voting on 50 randomly selected questions from the GSM8k test dataset, generating 5 solutions per question.

Our model	44%
Our model with majority vote	46% +4.5%

Conclusions and Further Directions:

Conclusions: We efficiently fine-tuned the Llama 3-8B model on a large dataset of math problems, achieving a substantial performance boost relative to the baseline. Despite this improvement, our model's performance still falls short of the latest state-of-the-art models. Specifically, while we observed notable gains in accuracy, especially on the GSM8k dataset, there remains significant room for improvement in solving more complex mathematical problems as evidenced by our results on the MATH dataset. Further refinement and advanced techniques are necessary to bridge the gap and achieve superior performance in mathematical reasoning tasks.

Further Directions:

- Training a larger base model (e.g. Llama 3-70B)
- Training for two to three epochs on a larger math problem dataset.
- Training an LLM as a verifier to judge the correctness of model-generated solutions, aiming to reduce false positives and improve overall accuracy.

References:

1. https://huggingface.co/unsloth/llama-3-8b-bnb-4bit
2. https://github.com/Xwin-LM/Xwin-LM/tree/main/Xwin-Math
3. https://huggingface.co/datasets/meta-math/MetaMathQA
4. https://github.com/openai/grade-school-math
5. https://github.com/hendrycks/math